第41卷，第11期
2021年11月

光 谱 学 与 光 谱 分 析
Spectroscopy and Spectral Analysis

Vol. 41, No. 11, pp3357-3362
November, 2021

基于太赫兹光谱技术的贝母品种鉴别方法研究

刘燕德，徐 振，胡 军，李茂鹏，崔惠桢

华东交通大学机电与车辆工程学院，江西 南昌 330013

摘 要 贝母是广泛应用于临床实践的中药材，其中川贝母尤为珍贵，存在掺假及伪冒现象，伪劣贝母会对用药者的健康产生不良影响。太赫兹时域光谱(Terahertz time domain spectroscopy)具有瞬态性、宽带性、安全性和穿透性等许多优越特性，近年来在药食无损检测领域十分活跃。以四种常见贝母(川贝母、平贝母、伊贝母、浙贝母)为研究对象，探究利用太赫兹时域光谱技术鉴别贝母品种的可行性。利用TAS7500TS太赫兹光谱系统采集贝母样品在0.6～3.0 THz范围内的光谱，并结合化学计量学方法进行预处理与建立分类模型。当分类数量为二时，称为二分类问题，当分类数量超过二时称为多分类问题。利用偏最小二乘判别分析(PLS-DA)建立四种贝母的二分类模型；使用Savitzky-Golay(S-G)平滑、多元散射校正(MSC)、标准正态变量变换(SNV)、移动平均、基线偏移校正(Baseline offset)对原始光谱进行预处理，再采用主成分分析(PCA)对预处理后的数据进行降维，以减少数据运算量、简化运算，最后建立随机森林(RF)、支持向量机(SVM)、反向传播神经网络(BPNN)多分类模型。结果显示：川-伊贝母二分类鉴别模型正确率为93.333%，平-浙贝母二分类鉴别模型正确率为98.333%，其他四种二分类鉴别模型正确率均为100%。对建立的多分类模型进行对比分析发现SVM结合SNV建模效果最好，其中川贝母正确率为95.349%，伊贝母正确率为96.552%，平贝母与浙贝母正确率均为100%，整体正确率高达97.490%。研究结果表明利用太赫兹时域光谱技术鉴别不同品种贝母是可行的，并建立了分类效果较好的SNV-SVM多分类模型，为把控中药材质量提供一种新的手段，对维护中药材市场的正常运转具有重要的意义。

关键词 太赫兹光谱技术；贝母；二分类；多分类

中图分类号：O434.3 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2021)11-3357-06

引 言

贝母为多年生草本植物，其鳞茎部分常作药用。《本草经集注》说："形似聚贝子"，名为贝母，主治止咳化痰、清热散结等。常将贝母类药材分为：川贝母、浙贝母、平贝母、伊贝母、土贝母等，贝母品种不同药效也略有差异，且极易混淆。川贝母是贝母中的珍品，药用效果相对较高，生存环境脆弱，生长周期漫长，产量相对较低，市场需求较高，价格极高，易被冒充[1]。广大群众鉴别易混淆中草药多基于传统"一看、二闻、三尝"的经验鉴别，此方法需要积累丰富的经验，且极易出错、难以鉴别高仿。近年来基于理化分析的高效液相色谱法(high PerforMance liquid chromatography，HPLC)、薄层色谱法(thin layer chromatography，TLC)、气相色谱法(gas chromatography，GC)、质谱法(mass spectrometry，MS)以及联用技术等方法也被广泛应用于测定中草药的主要成分及鉴别种类[2-3]。但此类检测手段需要复杂的样品处理，以及专业人员的操作，费时费力[4]。因此有必要开发探索一种新的检测手段来弥补传统方法的缺陷。

太赫兹波频率处于0.1～10 THz之间，具有能量低、频谱宽、穿透强与吸收强的特征，基于太赫兹光谱的鉴别检测技术具有识别率高、耗时短、操作简单等优势，是一种新颖的检测手段。太赫兹光谱独特的优势，使其近些年在食品、生物、化工、材料和医药检测等领域得到广泛应用。中草药的药用成分结构复杂，其有机分子之间的弱相互作用和振动跃迁以及晶体中的低频振动和吸收频率大多数处于太赫兹波段范围内。这些振动充分反映了中草药的分子结构及相关信息，因此使得太赫兹光谱技术对中药材检测鉴别成为可能。

收稿日期：2020-09-30，**修订日期**：2021-01-19

基金项目：国家"十二五"(863)计划项目(SS2012AA101906)，南方山地果园智能化管理技术与装备协同创新中心项目(赣教高字[2014]60号)，国家自然科学基金项目(31760344)，江西省教育厅科学技术研究青年项目(GJJ190348)，江西省博士研究生创新资金项目(YC2019-B106)资助

作者简介：刘燕德，女，1967年生，华东交通大学机电与车辆工程学院教授 e-mail：jxliuyd@163.com

马品等[5]使用太赫兹光谱检测技术对天麻含水量进行检测，表明太赫兹可以在测定中药饮片含水量中得到应用。徐哲等[6]为对五种不同产地、不同批次的鸡血藤和大血藤进行鉴别，采用太赫兹光谱技术结合光谱角算法对鸡血藤与大血藤进行分类，效果较为理想，两类中药样品总计 100 组数据的分类正确率达到 95%。Zhang 等[7]先后对中药中的添加剂、易混淆中草药、有毒中草药进行鉴别研究，效果均较为理想。李辰等[8]对正品与伪劣冬虫夏草进行鉴别，发现冬虫夏草正品存在 1.01 THz 和 1.13 THz 特征吸收峰，根据吸收峰实现对正伪冬虫夏草的鉴别。杨少壮等[9]对陈皮的 THz 图谱进行分析以判断储存年份，建立了基于主成分分析-支持向量机(PCA-SVM)的高效陈皮贮存年限预测模型，其年限预测准确度可达 94%以上。上述研究利用太赫兹光谱技术从不同的角度对中药材的品质进行把控，为后续研究者提供了经验借鉴。本研究将太赫兹光谱技术与多种化学计量学方法结合，对川贝母、平贝母、伊贝母、浙贝母四种不同品种的贝母进行定性鉴别，试图探索一种快速无损的贝母品种鉴别方法。

1 实验部分

1.1 仪器

实验所用的 THz-TDS 系统由日本 Advantest 公司研制，系统使用两个超短脉冲激光器(1.55 μm)分别作为偏置输出(太赫兹波产生)和信号输入(太赫兹波探测)的光源。飞秒激光脉冲输出功率 20 mW，中心波长 1 550 nm，脉宽 50 fs，重复频率 50 MHz。由于太赫兹波对水分比较敏感，为减少实验误差，将太赫兹电磁辐射通过的光路封闭在干燥箱内，并通入干燥空气，在实验过程中，湿度保持在 10%的恒定值，温度 25 ℃。图 1 为实验所用设备的原理图。

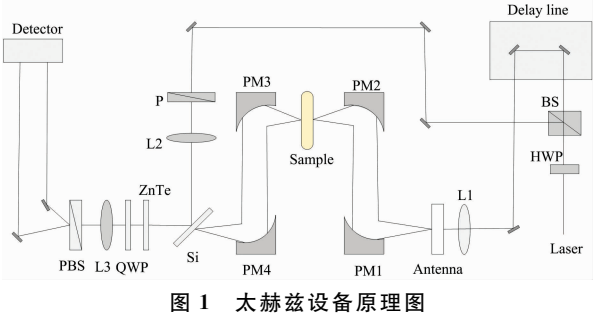

图 1 太赫兹设备原理图

Fig. 1 Schematic diagram of Terahertz equipment

1.2 样品制备

实验所用 4 种贝母均采购于中药房，首先将四种贝母样品放入干燥箱中 50 ℃，干燥 2 h，磨成粉末再过 200 目筛，密封保存。每种样品均按照同一比例(62.5%)加入高密度聚乙烯，用涡旋振荡器震荡 3 min，以确保聚乙烯与样品粉末充分混匀。压片时每次称取(0.1±0.005) g，设置压力 16 MPa，压片控制时长 2 min，使每个样品片厚度维持在 0.8 mm 左右，放入密封袋保存。四种样品各压制 25 个待测样品片，共计 100 个。每个样品采集 5 个点，每个点采集 2 次。为保证采集环境的稳定性，将样品放入样品仓后，等待 3 min 后开始采集光谱，每类贝母的太赫兹时域光谱各 250 条，共采集到 1 000 条光谱。

1.3 数据采集

所有测量均采用图 1 所示的 THz-TDS 系统进行。依据 Dorney 等[10]和 Dragoman 等[11]提出的光学参数提取方法提取所需的光谱信息，参数包括透射率、折射率、吸收系数等，此类参数对具有厚度均匀且两面平行固体样品，在透射模式下的太赫兹光谱吸收特性进行描述。实验记录参考太赫兹时域信号 $E_{ref}(t)$和样本的太赫兹时域信号 $E_{sam}(t)$，利用快速傅里叶变换(fast Fourier transform，FFT)算法可以得到光谱。根据菲涅耳公式，大多数低损耗材料的 THz 振幅透射率 T 可以表示为

$$T(\omega)=\frac{E_{sam}(\omega)}{E_{ref}(\omega)}=A\exp(-i\varphi)\approx \frac{4n}{(1+n)^2}\exp\left[\frac{i\omega(N-1)d}{c}\right] \tag{1}$$

式(1)中，$E_{ref}(\omega)$和 $E_{sam}(\omega)$分别为入射和透射的 THz 频域谱；A 和 φ 分别为基准信号和样本信号的幅值比和相位差；$N=N+ik$ 为样品的复折射率，k 为消光系数；d 为试样厚度；ω 是角频率，c 是真空中的光速。由式(2)和式(3)得到折射率 $n(\omega)$和吸收系数 $\alpha(\omega)$。

$$n(\omega)=\frac{\varphi(\omega)c}{\omega d}+1 \tag{2}$$

$$\alpha(\omega)=\frac{2k(\omega)\omega}{c}=\frac{2}{d}=\ln\frac{4n(\omega)}{A(\varphi)(n(\omega)+1)^2} \tag{3}$$

1.4 数据处理流程

获取到的太赫兹光谱除包含其自身的物理化学信息外，还夹杂其他干扰信息，因此在使用化学计量学方法建模前，需要对原始光谱进行预处理，去除噪声。同时由于样品光谱数据量较大，还需要进行降维处理。采用 K-S 算法将光谱数据按 3∶1 随机分为建模集和预测集，分别建立二分类和多分类模型。图 2 为实验具体过程图。

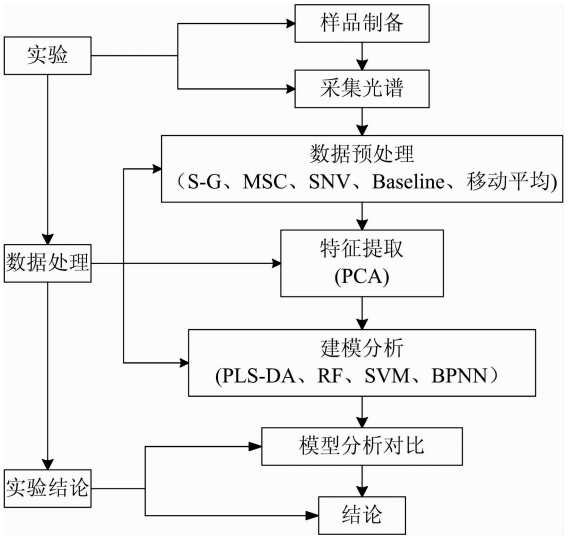

图 2 贝母分类流程图

Fig. 2 Flowchart of fritillary classification

1.5　算法介绍

预处理：实验为寻求对贝母样品光谱最优的预处理方法，主要用到移动平滑、S-G 平滑、多元散射校正(multiplicative scatter correction，MSC)、标准正态变量变换(standard normal variable transformations，SNV)和基线偏移校正(Baseline offset)五种预处理方法，进行光谱预处理是为了消除光谱的冗余信息，提高模型稳定性与准确性。

主成分分析(principal component analysis，PCA)是常用在光谱分析中进行数据降维，以减少数据运算量，其基本原理是通过正交变换将相关变量转换为线性不相关的变量，经过变换之后得到原始光谱的主成分，同时这些主成分基本能够代替原始光谱的信息[12]。累计方差贡献率决定主成分的个数，当累计方差贡献率能够提供原始变量的绝大部分信息时，即根据方差贡献率与主成分数关系图确定所需的主成分数。

偏最小二乘判别分析(partial least squares-discriminant analysis，PLS-DA)是一种基于偏最小二乘(PLS)的多变量分析方法，该方法将主成分分析与相关性分析结合，对光谱数据与分类变量进行线性拟合[13]。

随机森林(random forest，RF)是基于决策树的一种机器学习方法，其与自然界中由树组成森林的概念类似，以决策树作为基本组成单元，决策树之间彼此独立。根据若干个有差异性的样本子集建立决策树，再采用投票机制得到最终判断。由于其具有优秀的预测精度和较小的运算量，随机森林目前已经得到广泛的关注[14]。

支持向量机(support vector machine，SVM)是一种基于结构风险最小化准则的模式识别方法，该方法对小样本、非线性和高维度问题中优势显著。本实验主要采用高斯核函数的 SVM 分类，此方法需要寻求惩罚因子 C 和核函数 g 两个参数的最佳优化值，两个参数对分类效果有着重要影响。

反向误差传输神经网络(back propagation neural network，BPNN)是一种前馈多层神经网络，由非线性变换神经单元组成，可以实现输入和输出间的任意非线性映射，非线性映射逼近能力和泛化能力强大，在建立大样本的非线性校正模型中被广泛应用[15]。

2　结果与讨论

2.1　各类贝母的 THz 光谱

图 3 为四种贝母在 0.6～3.0 THz 波段的平均吸收曲线，未对光谱做任何预处理。可以看出四种样品的光谱曲线趋势较为相似，均无明显的吸收峰，在低频区域四种样品的平均光谱曲线重叠较为严重：在高频区域川贝母的吸收系数明显低于其他三类贝母，且平贝母、伊贝母、浙贝母区分不够明显，这可能是由于四种贝母某些药用成分含量不同造成的。

2.2　建模与分析

2.2.1　贝母样品的二分类鉴别

根据采集到的样品原始光谱，建立 PLS-DA 定性分析模型对川贝母与其他三类贝母进行鉴别区分。每两种样品光谱数据各 250 组，共计 500 组样品数据，随机选取 120 组作为验证集，380 组光谱数据为建模集。表 1 为二分类 PLS-DA 判别模型正确率。

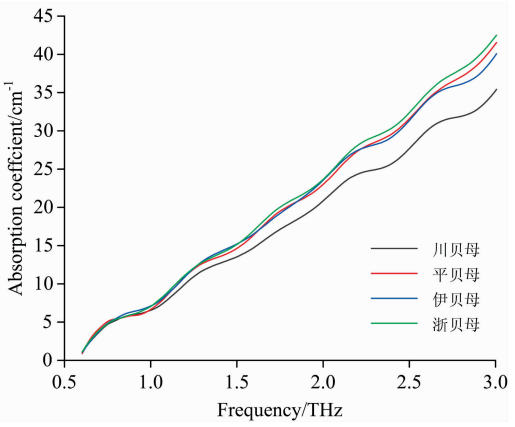

图 3　四种贝母的平均吸收光谱

Fig. 3　Mean absorption spectra of the four fritillaria species

表 1　PLS-DA 模型分类正确率

Table 1　Classification accuracy of PLS-DA model

模型	整体正确率/%	各类正确率/%
川贝母	100	100
平贝母		100
川贝母	93.333	91.667
伊贝母		95
川贝母	100	100
浙贝母		100
平贝母	100	100
伊贝母		100
平贝母	98.333	96.923
浙贝母		100
伊贝母	100	100
浙贝母		100

共建立了 6 个二分类模型，其中川贝母-平贝母、川贝母-浙贝母、平贝母-伊贝母、伊贝母-浙贝母 4 个二分类模型正确率均为 100%。川贝母-伊贝母二分类模型的整体正确率为 93.333%，其中川贝母的正确分类率为 91.667%，伊贝母的正确分类率为 95%。平贝母-浙贝母二分类模型的整体正确率为 98.333%，其中平贝母的正确分类率为 96.923%，浙贝母的正确分类率为 100%。二分类模型整体分类效果较好。

根据原始光谱数据建立 PLS-DA 模型进行预测时，其中川贝母-伊贝母鉴别时，5 个川贝母被错误识别成伊贝母，3 个伊贝母被错误识别成川贝母。进行平贝母-浙贝母鉴别时，2 个平贝母被错误识别成浙贝母，其他贝母均无错分现象，整体效果较好。图 4 为各种 PLS-DA 二分类模型。

2.2.2　贝母样品的多分类鉴别

当多种贝母掺杂在一起时，采用 PLS-DA 鉴别，结果精度较差，为寻求最佳解决贝母的多分类问题，在采用多种预处理方法多光谱数据进行预处理之后，利用主成分分析提取

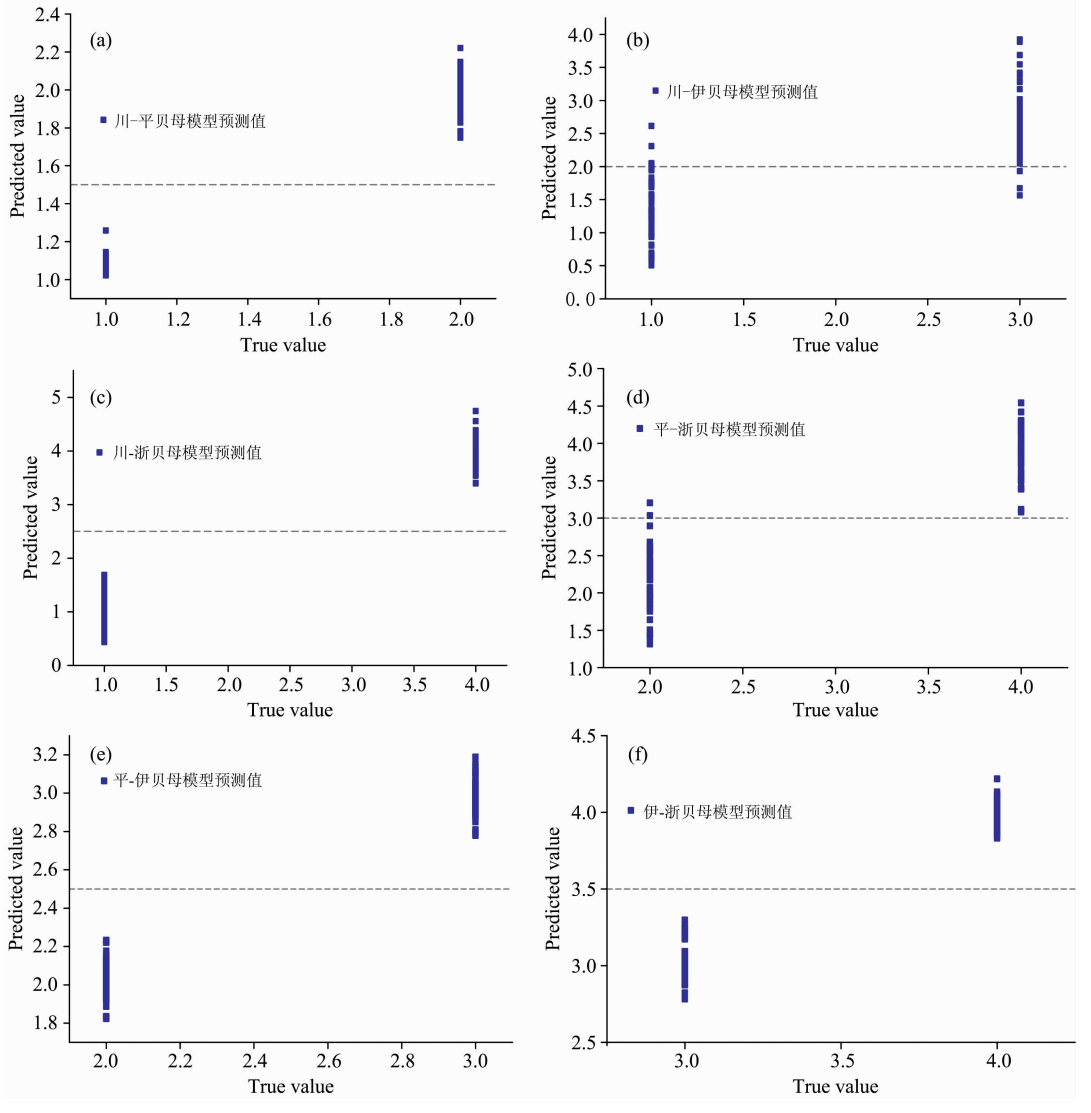

图 4 六种 PLS-DA 二分类模型

Fig. 4 PLS-DA dichotomy model of 6 categories

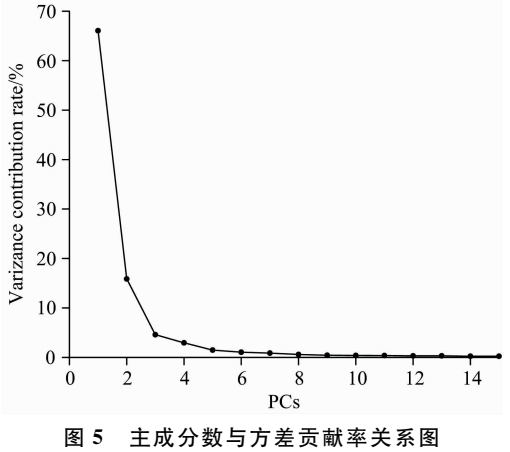

图 5 主成分数与方差贡献率关系图

Fig. 5 Relationship between principal component number and variance contribution rate

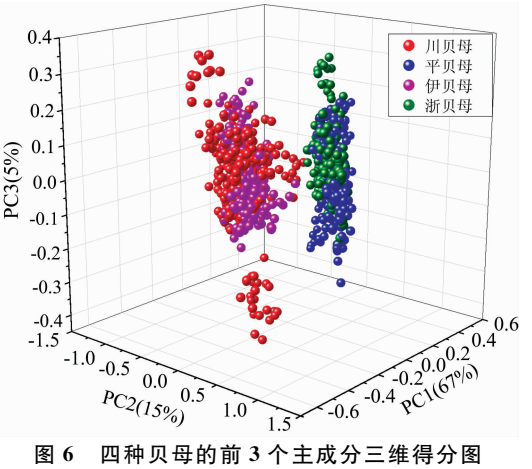

图 6 四种贝母的前 3 个主成分三维得分图

Fig. 6 Three-dimensional scores of first three principal components of four fritillaria species

数据的主要特征，降低光谱数据的维度。图 5 为四种贝母样品的太赫兹光谱经过 SNV 预处理之后的主成分数与方差贡献率关系图，图 6 为前三个主成分评分图。当主成分大于 13 时，随着主成分数的增加，方差贡献率增幅趋于 0，累积方差贡献率达到 95%。

由于贝母成分复杂，特征吸收峰不明显，无法通过直接观察进行分类，需借助机器学习算法，故在对主成分分析之后的数据分别建立 RF，SVM 和 BPNN 多分类模型。表 2 为各模型鉴别四种贝母产地的具体正确率。

表 2　贝母多分类结果对比

Table 2　Comparison of multiple classification results of fritillaria

分类模型	预处理方法	主成分数	川贝母/%	平贝母/%	伊贝母/%	浙贝母/%	整体/%
RF	无	6	91.228	95.000	84.746	93.750	91.250
	S-G	5	80.851	90.045	93.750	90.323	89.583
	MSC	13	87.273	100.000	85.714	100.000	94.167
	SNV	13	86.441	98.508	96.364	100.000	95.417
	移动平均	4	92.452	86.793	90.000	90.625	90.000
	Baseline offset	5	87.500	98.413	94.915	88.708	95.200
SVM	无	6	90.625	78.333	90.000	96.868	88.750
	S-G	5	80.090	72.131	84.746	90.566	82.083
	MSC	13	91.667	100.000	96.491	100.000	96.250
	SNV	13	97.490	100.000	96.552	100.000	97.490
	移动平均	4	82.090	62.500	89.474	80.769	78.333
	Baseline offset	5	83.333	94.444	90.000	83.333	87.500
BPNN	无	6	79.6880	83.333	74.067	80.357	79.167
	S-G	5	68.657	55.738	72.881	62.264	65.000
	MSC	13	61.857	80.000	85.965	88.636	78.750
	SNV	13	82.558	83.333	74.1380	73.810	79.167
	移动平均	4	76.119	40.625	78.944	48.077	61.250
	Baseline offset	5	72.222	77.778	58.333	74.074	70.417

其中 BPNN 类模型的效果最差，尽管结合多种预处理方法，但整体正确率均未超过 80%。可能是由于 BPNN 可以对大量数据进行模型训练，但其极易陷入模型训练速度较慢的状态。RF 结合 SNV 建模时，效果较好，正确率为 95.417%，共计 11 个贝母样品被错误分类。综合三类模型，SVM 结合 SNV 建模效果最好，整体正确率高达 97.490%，预测集剔除一个异常点之后共计 239 个样本，被错误分类 6 列，其中川贝母 4 例，正确率为 97.490%，浙贝母 1 例，正确率为 96.552%，平贝母与伊贝母均无出错。

3　结　论

以川贝母、平贝母、伊贝母、浙贝母四种贝母为例，介绍太赫兹时域光谱技术结合化学计量学方法在中药材定性鉴别中的应用。对原始光谱预处理之后，采用主成分分析(PCA)提取主要特征，再建立二分类判别模型，其中川贝母-伊贝母二分类模型正确率为 93.333%，平贝母-浙贝母二分类模型正确率为 98.333%，其他二分类模型正确率均为 100%，表明 PLS-DA 可以实现贝母的两两准确分类；最后分别建立随机森林(RF)、支持向量机(SVM)、反向误差神经网络(BPNN)建立多分类模型并进行对比，SVM 结合 SNV 预处理建模效果最好，整体正确率高达 97.490%。这表明四种贝母样品的太赫兹吸收光谱虽均无明显的吸收峰，但经过光谱预处理结合合理的分类模型，可以实现相似贝母的准确区分。本研究对维护中药材的安全以及中国传统医药市场秩序具有重要的意义，也为后期利用太赫兹时域光谱技术对中药材更深层次的研究提供理论借鉴。

References

[1] YANG Jian, LI Jing, XUE Wei-na, et al(杨　健，李　靖，薛维娜，等). Chinese Traditional Patent Medicine(中成药)，2020，42(5)：1262.

[2] WANG Hong-wei, FANG Bo, ZHANG Lei, et al(王宏伟，方　波，张　磊，等). Chinese Traditional Patent Medicine(中成药)，2020，42(4)：986.

[3] LI Ran, LI Jing, TONG Qiao-zhen, et al(李　然，李　静，童巧珍，等). Journal of Guangdong Pharmaceutical University(广东药科大学学报)，2019，35(5)：624.

[4] Li Rong, Zeng Canbiao, Li Junni, et al. Analytical Letters，2020，53(11)：1667.

[5] MA Pin, YANG Yu-ping(马 品, 杨玉平). Journal of Terahertz Science and Electronic Information Technology(太赫兹科学与电子信息学报), 2017, 15(1): 26.
[6] XU Zhe, HE Ming-xia, LI Peng-fei, et al(徐 哲, 何明霞, 李鹏飞, 等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2017, 37(1): 42.
[7] Zhang H, Li Z, Chen T, et al. Journal of Applied Spectroscopy, 2018, 85(1): 197.
[8] LI Chen, WEI Cheng-hao, WANG Zhi-qi, et al(李 辰, 魏丞昊, 王志琪, 等). Journal of Shenzhen University · Science & Engineering (深圳大学学报 · 理工版), 2019, 36(2): 213.
[9] YANG Shao-zhuang, LI Can, LI Chen, et al(杨少壮, 李 灿, 李 辰, 等). Modern Food Science & Technology(现代食品科技), 2019, 35(12): 258.
[10] Dorney T D, Baraniuk R G, Mittleman D M. J. Opt. Soc. Am. A, 2001, 18(7): 1562.
[11] Dragoman D, Dragoman M. Appl. Optics, 2004, 43(19): 3848.
[12] WANG Cheng, SHI Ji-yi, ZHENG Gang, et al(王 成, 史继毅, 郑 刚, 等). Optical Instruments(光学仪器), 2020, 42(2): 26.
[13] Liu Jianjun, Mao Lili, Ku Jingfeng, et al. Optik, 2017, 142: 483.
[14] Poona N K, Van Niekerk A, Nadel R L, et al. Applied Spectroscopy, 2016, 70(2): 322.
[15] HU Jun, LIU Yan-de, SUN Xu-dong, et al(胡 军, 刘燕德, 孙旭东, 等). Laser & Optoelectronics Progress(激光与光电子学进展), 2020, 57(7): 073002.

Research on Variety Identification of Fritillaria Based on Terahertz Spectroscopy

LIU Yan-de, XU Zhen, HU Jun, LI Mao-peng, CUI Hui-zhen
School of Mechanical, Electrical and Vehicle Engineering, East China Jiaotong University, Nanchang 330013, China

Abstract Fritillary is widely used in clinical practice of Chinese medicinal materials, especially Fritillaria cirrhosa Don. There are adulteration and fake phenomenon, fake fritillary will have a negative impact on the health of the drug users. Terahertz Time-Domain spectroscopy has many advantages of transient, broadband, safety, penetration, etc. In recent years, Terahertz Time-Domain spectroscopy is very active in drug and food non-destructive detection. In this experiment, four common fritillaria species (Fritillaria cirrhosa Don, Fritillaria ussuriensis Maxim, Fritillaria pallidiflora Schrenk, and Fritillaria thunbergii) were taken as the research objects to explore the feasibility of using terahertz time-domain spectroscopy to identify fritillaria species. In this experiment, the TAS7500TS Terahertz spectrum system was used to collect the spectra of fritillate samples in the range of 0.6～3.0 THz, and the stoichiometric method was combined for pretreatment and classification model establishment. When the number of categories is 2, it is called Binary classification; when the number of categories exceeds 2, it is called Multiple classifications. Four kinds of fritillary were established by Partial Least Squares Discriminant Analysis (PLS-DA). Initial spectra are treated with Savitzky-Golay (S-G) smoothing, Multiplicative Scatter (MSC) Correction, Standard Normal Variable Transformations, moving averages, or Baseline. Principal Component Analysis is performed. PCA can reduce the dimensionality of the preprocessed data to reduce the amount of data computation and simplify the operation. Finally, a multi-classification model of Random Forest (RF), Support Vector Machine (SVM) and Back Propagation Neural Network (BPNN) can be established. The discriminant accuracy rate of the model was 93.333% for Fritillaria cirrhosa Don-Fritillaria pallidiflora Schrenk, 98.333% for Fritillaria cirrhosa Don-Fritillaria thunbergii, and 100% for all the other four biocalcification models. The accuracy of the other four dichotomies was 100%. By comparing and analyzing the established multi-classification models, it was found that the SVM combining SNV modeling effect is best, the Fritillaria cirrhosa Don accuracy is 95.349%, the Fritillaria pallidiflora Schrenk accuracy is 96.552%, the accuracy rate of Fritillaria ussuriensis Maxim and Fritillaria thunbergii was 100%. The overall accuracy rate was up to 97.490%. This research shows that it is feasible to use Terahertz Time-Domain spectroscopy to identify different fritillaria varieties, and a SNV-SVM multi-classification model with good classification effect is established, which provides a new means to control the quality of traditional Chinese medicine and is of great significance to maintain the normal operation of the traditional Chinese medicine market.

Keywords Terahertz spectroscopy; Fritillaria; Binary classification; Multiple classification

(Received Sep. 30, 2020; accepted Jan. 19, 2021)